A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics

Christin, Christin; Hoefsloot, Huub C. J.; Smilde, Age K.; Hoekman, B.; Suits, Frank; Bischoff, Rainer; Horvatovich, Peter

Published in:
Molecular & Cellular Proteomics

DOI:
10.1074/mcp.M112.022566

Publication date:
2013

A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics*[S]

Christin Christin‡§, Huub C. J. Hoefsloot§¶, Age K. Smilde§¶, B. Hoekman‡§, Frank Suits‖, Rainer Bischoff‡§, and Peter Horvatovich‡§**

In this paper, we compare the performance of six different feature selection methods for LC-MS-based proteomics and metabolomics biomarker discovery—*t* test, the Mann–Whitney–Wilcoxon test (*mww* test), nearest shrunken centroid (NSC), linear support vector machine–recursive features elimination (SVM-RFE), principal component discriminant analysis (PCDA), and partial least squares discriminant analysis (PLSDA)—using human urine and porcine cerebrospinal fluid samples that were spiked with a range of peptides at different concentration levels. The ideal feature selection method should select the complete list of discriminating features that are related to the spiked peptides without selecting unrelated features. Whereas many studies have to rely on classification error to judge the reliability of the selected biomarker candidates, we assessed the accuracy of selection directly from the list of spiked peptides. The feature selection methods were applied to data sets with different sample sizes and extents of sample class separation determined by the concentration level of spiked compounds. For each feature selection method and data set, the performance for selecting a set of features related to spiked compounds was assessed using the harmonic mean of the recall and the precision (*f*-score) and the geometric mean of the recall and the true negative rate (*g*-score). We conclude that the univariate *t* test and the *mww* test with multiple testing corrections are not applicable to data sets with small sample sizes ($n = 6$), but their performance improves markedly with increasing sample size up to a point ($n > 12$) at which they outperform the other methods. PCDA and PLSDA select small feature sets with high precision but miss many true positive features related to the spiked peptides. NSC strikes a reasonable compromise between recall and precision for all data sets independent of spiking level and number of samples. Linear SVM-RFE performs poorly for selecting features related to the spiked compounds, even though the classification error is relatively low. ***Molecular & Cellular Proteomics 12: 10.1074/mcp.M112.022566, 263–276, 2013.***

Biomarkers play an important role in advancing medical research through the early diagnosis of disease and prognosis of treatment interventions (1, 2). Biomarkers may be proteins, peptides, or metabolites, as well as mRNAs or other kinds of nucleic acids (*e.g.* microRNAs) whose levels change in relation to the stage of a given disease and which may be used to accurately assign the disease stage of a patient. The accurate selection of biomarker candidates is crucial, because it determines the outcome of further validation studies and the ultimate success of efforts to develop diagnostic and prognostic assays with high specificity and sensitivity. The success of biomarker discovery depends on several factors: consistent and reproducible phenotyping of the individuals from whom biological samples are obtained; the quality of the analytical methodology, which in turn determines the quality of the collected data; the accuracy of the computational methods used to extract quantitative and molecular identity information to define the biomarker candidates from raw analytical data; and finally the performance of the applied statistical methods in the selection of a limited list of compounds with the potential to discriminate between predefined classes of samples. *De novo* biomarker research consists of a biomarker discovery part and a biomarker validation part (3). Biomarker discovery uses analytical techniques that try to measure as many compounds as possible in a relatively low number of samples. The goal of subsequent data preprocessing and statistical analysis is to select a limited number of candidates, which are subsequently subjected to targeted analyses in large number of samples for validation.

Advanced technology, such as high-performance liquid chromatography–mass spectrometry (LC-MS),[1] is increas-

From ‡Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands; §Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands; ¶Swammerdam Institute for Life Sciences, University of Amsterdam, 1090 GE Amsterdam, The Netherlands; ‖IBM T. J. Watson Research Centre, Yorktown Heights, New York 10598

Received July 24, 2012, and in revised form, September 30, 2012
Published, MCP Papers in Press, October 31, 2012, DOI 10.1074/mcp.M112.022566

[1] The abbreviations used are: CSF, cerebrospinal fluid; IQR, interquartile range; LC-MS, liquid chromatography–mass spectrometry; NSC, nearest shrunken centroid; PCDA, principal component discriminant analysis; PLSDA, partial least squares discriminant analysis; SVM-RFE, support vector machine–recursive features elimination; *mww* test, Mann–Whitney–Wilcoxon test.

ingly applied in biomarker discovery research. Such analyses detect tens of thousands of compounds, as well as background-related signals, in a single biological sample, generating enormous amounts of multivariate data. Data preprocessing workflows reduce data complexity considerably by trying to extract only the information related to compounds resulting in a quantitative feature matrix, in which rows and columns correspond to samples and extracted features, respectively, or vice versa. Features may also be related to data preprocessing artifacts, and the ratio of such erroneous features to compound-related features depends on the performance of the data preprocessing workflow (4). Preprocessed LC-MS data sets contain a large number of features relative to the sample size. These features are characterized by their *m/z* value and retention time, and in the ideal case they can be combined and linked to compound identities such as metabolites, peptides, and proteins. In LC-MS-based proteomics and metabolomics studies, sample analysis is so time consuming that it is practically impossible to increase the number of samples to a level that balances the number of features in a data set. Therefore, the success of biomarker discovery depends on powerful feature selection methods that can deal with a low sample size and a high number of features. Because of the unfavorable statistical situation and the risk of overfitting the data, it is ultimately pivotal to validate the selected biomarker candidates in a larger set of independent samples, preferably in a double-blinded fashion, using targeted analytical methods (1).

Biomarker selection is often based on classification methods that are preceded by feature selection methods (filters) or which have built-in feature selection modules (wrappers and embedded methods) that can be used to select a list of compounds/peaks/features that provide the best classification performance for predefined sample groups (*e.g.* healthy *versus* diseased) (5). Classification methods are able to classify an unknown sample into a predefined sample class. Univariate feature selection methods such as filters (*t* test or Wilcoxon–Mann–Whitney tests) cannot be used for sample classification. Other classification methods such as the nearest shrunken centroid method have intrinsic feature selection ability, whereas other classification methods such as principal component discriminant analysis (PCDA) and partial least squares regression coupled with discriminant analysis (PLSDA) should be augmented with a feature selection method. There are classifiers having no feature selection option that perform the classification using all variables, such as support vector machines that use non-linear kernels (6). Classification methods without the ability to select features cannot be used for biomarker discovery, because these methods aim to classify samples into predefined classes but cannot identify the limited number of variables (features or compounds) that form the basis of the classification (6, 7). Different statistical methods with feature selection have been developed according to the complexity of the analyzed data, and these have been extensively reviewed (5, 6, 8, 9). Ways of optimizing such methods to improve sensitivity and specificity are a major topic in current biomarker discovery research and in the many "omics-related" research areas (6, 10, 11). Comparisons of classification methods with respect to their classification and learning performance have been initiated. Van der Walt *et al.* (12) focused on finding the most accurate classifiers for simulated data sets with sample sizes ranging from 20 to 100. Rubingh *et al.* (13) compared the influence of sample size in an LC-MS metabolomics data set on the performance of three different statistical validation tools: cross validation, jack-knifing model parameters, and a permutation test. That study concluded that for small sample sets, the outcome of these validation methods is influenced strongly by individual samples and therefore cannot be trusted, and the validation tool cannot be used to indicate problems due to sample size or the representativeness of sampling. This implies that reducing the dimensionality of the feature space is critical when approaching a classification problem in which the number of features exceeds the number of samples by a large margin. Dimensionality reduction retains a smaller set of features to bring the feature space in line with the sample size and thus allow the application of classification methods that perform with acceptable accuracy only when the sample size and the feature size are similar.

In this study we compared different classification methods focusing on feature selection in two types of spiked LC-MS data sets that mimic the situation of a biomarker discovery study. Our results provide guidelines for researchers who will engage in biomarker discovery or other differential profiling "omics" studies with respect to sample size and selecting the most appropriate feature selection method for a given data set. We evaluated the following approaches: univariate *t* test and Mann–Whitney–Wilcoxon test (*mww* test) with multiple testing correction (14), nearest shrunken centroid (NSC) (15, 16), support vector machine–recursive features elimination (SVM-RFE) (17), PLSDA (18), and PCDA (19). PCDA and PLSDA were combined with the rank-product as a feature selection criterion (20). These methods were evaluated with data sets having three characteristics: different biological background, varying sample size, and varying within- and between-class variability of the added compounds. Data were acquired via LC-MS from human urine and porcine cerebrospinal fluid (CSF) samples that were spiked with a set of known peptides (true positives) at different concentration levels. These samples were then combined in two classes containing peptides spiked at low and high concentration levels. The performance of the classification methods with feature selection was measured based on their ability to select features that were related to the spiked peptides. Because true positives were known in our data set, we compared performance based on the *f*-score (the harmonic mean of precision and recall) and the *g*-score (the geometric mean of accuracy).

TABLE I

Description of the sample groups that were combined to give data sets 0a–b, 1a–c and 2a–c. This scheme was used to select files for the 100 repetitions of each combination of feature selection methods and data sets (see supplemental Table S2 for results). One pooled sample was used to prepare all spiked human urine samples (data sets 0a, 1a–c, and 2a–c) and one porcine CSF sample was used to prepare all spiked and non-spiked CSF samples (data set 0b)

Data set	Between- and within-class variability		Sample size per class
Data Set 0a Human urine	Low within-class variability High between-class variability High spike class = group A Low spike class = group H	0a	5 samples: all samples belongs to groups A (class 1) or class H (class 0)
Data Set 0b Porcine CSF	Low within-class variability High between-class variability High spike class = spiked samples Low spike class = non-spiked samples	0b	5 samples: all samples belongs to spiked samples (class 1) or non-spiked samples (class 0)
Data Set 1a–c Human urine	High within-class variability High between-class variability High spiked class = combination of groups A–C Low spike class = combination of groups F–H	1a 1b 1c	6 samples: two samples were randomly taken from each of the groups A–C (class 1) and F–H (class 0) 12 samples: four samples were randomly taken from each of the groups A–C (class 1) and F–H (class 0) 15: all samples from groups A–C (class 1) and F–H (class 0)
Data Set 2a–c Human urine	High within-class variability Low between-class variability High spiked class = combination of groups B–D Low spike class = combination of groups E–G	2a 2b 2c	6 samples: two samples were randomly taken from each of the groups B–D (class 1) and E–G (class 0) 12 samples: four samples were randomly taken from each of the groups B–D (class 1) and E–G (class 0) 15 samples: all samples from groups B–D (class 1) and E–G (class 0)

EXPERIMENTAL PROCEDURES

Data Set Design

Spiked Human Urine Data Set—Fifty urine samples were obtained from 15 healthy females and 35 healthy males over the age range of 26.9 to 72.9 years. Two hundred microliters were taken from each sample, creating one pooled urine sample. This pooled urine sample was used as biological background. The pooled urine was spiked with a tryptic digest (V5111; Promega, Madison, WI) of bovine carbonic anhydrase (C3934, Uniprot entry: P00921; Sigma, Steinheim, Germany), as well as with seven synthetic peptides at eight different dilutions—6.25, 12.5, 25, 50, 100, 200, 400, and 2000 times dilution (called groups A–H, respectively)—of a stock solution containing 240 μM trypsin-digested carbonic anhydrase and the following concentrations (in μM) of the seven synthetic peptides: VYV, 83; YGGFL, 57; DRVYIHPF, 29; YPFPGPI, 46; YPFPG, 60; GYYPT, 54; and YGGWL, 57. At each concentration level, the sample was analyzed five times using an Agilent G2445A LC/MSD-Trap-SL ion trap mass spectrometer (Agilent Technologies, Santa Clara, California, United States), resulting in 40 LC-MS chromatograms. These chromatograms were pre-processed with a constant resolution of 0.1 amu using the threshold avoiding proteomics pipeline (TAPP) (21) covering peaks with *m/z* values of 280 to 1500 amu and retention times between 30 and 85 min, resulting in a final common peak list of 29,529 features, with 151 of those originating from the added peptides (supplemental Table S1*A*). Features are items provided by the data pre-processing pipelines and hold quantitative information about measured compounds. Features therefore can be related to peptide isotopologues, peptides, or proteins, but they can be related to data processing artifacts as well. Positive features are features selected by the feature selection method, and negative features are the non-selected ones. Positive and negative features are considered as true or false according to whether they correspond to spiked peptides (true) or not (false). Spiked peptides form the basis of the ground truth. It is therefore crucial to determine accurately all features in a pre-processed data set that correspond to spiked peptides. Details of the sample preparation, LC-MS data acquisition, and method of assignment of features related to spiked peptides are provided in the supplementary material and in Ref. 4. All subjects that participated in this study gave their oral and/or written informed consent. The study protocol was in agreement with local ethical standards and the Helsinki declaration of 1964, as revised in 2004.

Seven data sets were derived from the 40 chromatograms composed of two sample classes with low (class 0) and high (class 1) spiking levels. These sample classes were obtained by combining different spiking levels and sample sizes, which resulted in data sets of different between- and within-class variability of spiked peptides. Table I provides a description of the most important characteristics of the derived data sets, and Fig. 5 provides a graphical overview of the sample size and within- and between-class variability of spiked peptides of the various data sets.

Data set 0a has large between-class variability and a class sample size of 5. High spiked class 1 was prepared from samples of group A, and low spiked class (class 0) from samples of group H, leading to low

within-class variability corresponding to analytical variability. The six remaining data sets were combined from three different sample sizes (6, 12, and 15, denoted by a, b, and c in the data set indexation) and a high and low between-class variability (indicated as 1 and 2, respectively; see Table I for details). Samples from groups A–C were used for the high spiked class (class 1), and sample groups F–H were used for the low spiked class (class 0) for data sets with high between-class variability (data set 1). Samples from groups B–D were used for the high spiked class (class 1), and sample groups E–G were used for the low spiked class (class 0) for data sets with low within-class variability (data set 2). Samples were selected randomly for data sets 1a and 2a, with six samples per class, and data sets 1b and 2b, with 12 samples per class, from the respective spiking groups, and all samples were used for data sets 1c and 2c, with 15 samples per class. Feature selection for each combination of feature selection methods and data sets was repeated 100 times, each time using a different combination of samples in the inner and outer loop of the double cross validation procedure.

Spiked Porcine CSF Data Set—The collection, storage, sample preparation, and LC-MS analysis protocol of the tryptic digest of porcine CSF was described by Hoekman *et al.* (4) and Suits *et al.* (21). The spiked samples were prepared by mixing 20 μl CSF digest with 20 μl of a tryptic digest of horse heart cytochrome C (Fluka, part # 30396) at concentrations of 25 fmol/μl. Spiked and non-spiked samples were aliquoted in five tubes containing 8 μl each. Spiked and non-spiked trypsin-digested CSF was injected five times (4 μl from individual vials) in random order (the amount of injected spiked cytochrome C was 50 fmol) into an Agilent QTOF 6510 equipped with a chip interface. Raw LC-MS data was exported in mzData format using Quantitative Analysis (B.03.01) in the MassHunter software package in centroid mode to limit file size and analysis time. These data were processed using TAPP (21) in a manner similar to that used for the human urine dataset. After preprocessing, a total of 9889 features were detected in this data set, from which 38 corresponded to the spiked horse heart cytochrome C (see further details on *m/z* and rt of spiked-in features in supplemental Table S1*B*).

Data set 0b was created from non-spiked (class 0) and high spiking level (class 1) sample classes with a sample size of 5, and with low within-class variability corresponding to analytical variability and large between-class variability of spiked-in peptides similar to the human urine data set 0a.

Biomarker Discovery Methods

Univariate Tests—The parametric univariate *t* test ranks features according to their *p* value and is not a classification method. Because the data sets contained 6 to 15 samples per class, it was difficult to test the normality of the data, which in this case is the distribution of the peak intensities. We therefore also used a non-parametric univariate filter, the *mww* test. Because the data sets contained a large number of features, we corrected the calculated *p* values for multiple testing using the Benjamini–Hochberg approach (14). A feature was considered significant when the *p* value was below 0.05 after multiple testing corrections.

Semi-multivariate–NSC—The NSC approach aims to find a set of features that gives the minimum classification error or the highest sum of correct class probabilities in a set of training samples using double cross validation by progressively eliminating features that do not contribute to the construction of the shrunken class centroid. This method was proposed by Tibshirani *et al.* for the classification of cancer samples based on microarray data (15, 16). The double cross validation scheme for this method is outlined in Fig. 1. Other classification methods with feature selection used in this study were implemented according to similar double cross validation schemes (see supplemental Figs. S1 and S2).

The distance d_{ik} between a feature i in class k and its respective overall centroid is calculated as the difference between the within-class mean $\bar{x}_{ik}$ and the overall mean $\bar{x}_i$, normalized to the standard error. The standard error (Eq. 1) is calculated using the pooled within-class standard deviation of the respective feature s_i, a constant s_0 (median of the standard deviation s_i across all features) to avoid large distances due to small standard deviations, and the constant $x_k = \sqrt{1/n_k - 1/n}$. The shrinkage threshold Δ is iteratively subtracted from this distance, and features whose shrunken distance in all classes is zero or negative are eliminated. A test sample x^* is attributed to the class to which it has the highest class probability N_k. The discriminant score for class k and for test sample x^* is $\mathrm{N}_k(x^*)$, which is the sum of the standardized squared distances between each relevant feature in the test sample x^* and the k^{th} shrunken centroid $\bar{x}'_{ik}$ corrected by the prior probability π_k of class k (Eq. 2). This distance is basically similar to a simple diagonal covariance matrix between the test sample and the shrunken centroid of the respective class. Because feature elimination is done univariately but classification of the test sample to the class-specific shrunken centroid at a given shrinkage is calculated multivariately, we call this a semi-multivariate method.

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot (s_i - s_0)} \qquad \text{(Eq. 1)}$$

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \bar{x}'_{ik})^2}{s_i^2} - 2\log\pi_k \qquad \text{(Eq. 2)}$$

Based on the discriminating score, we can calculate the class probability $\hat{p}_k(x^*)$ (Eq. 3), which is the probability of sample x^* belonging to class k.

$$\hat{p}_k(x^*) = \frac{e^{-(1/2)\delta_k(x^*)}}{\sum_{l=1}^{K} e^{-(1/2)\delta_l(x^*)}} \qquad \text{(Eq. 3)}$$

Once the class probability has been calculated for each test sample and for each shrinkage value, the probability of the true class for each sample is summed up. We now have two measurements, based on which we select the optimum subset of features: (a) the subset that minimizes the classification error (Eq. 2), and (b) the subset that maximizes the sum of true class probabilities (Eq. 3). In our study, the optimum shrinkage was chosen based on the maximum true class probability of the test data set, because it gives a continuous plot and a well-defined optimal shrinkage value. Once the optimum shrinkage was obtained in the inner cross validation loop (see Fig. 1), it was applied to the training data set in the inner loop to obtain the optimal corresponding feature set. Based on this feature set, the discriminant and the true probability scores were calculated for the independent test data set in the outer loop to assess the classification model performance. Because each passage through the outer loop yields a different value for the optimum shrinkage, we calculated the median of the correct class probability scores at each shrinkage value from all the outer loops at the end of the double cross validation scheme. Fluctuation of the true class probability curves as a function of the optimal shrinkage values obtained for various outer loop evaluations reflects the stability of the model and small class differences in the studied data set. The shrinkage at the maximum of the median true class probability curve was used to select the optimal feature set using all samples in the data set.

Multivariate SVM-RFE—A support vector machine (SVM), originally proposed by Vapnik (22), is a multivariate supervised learning method

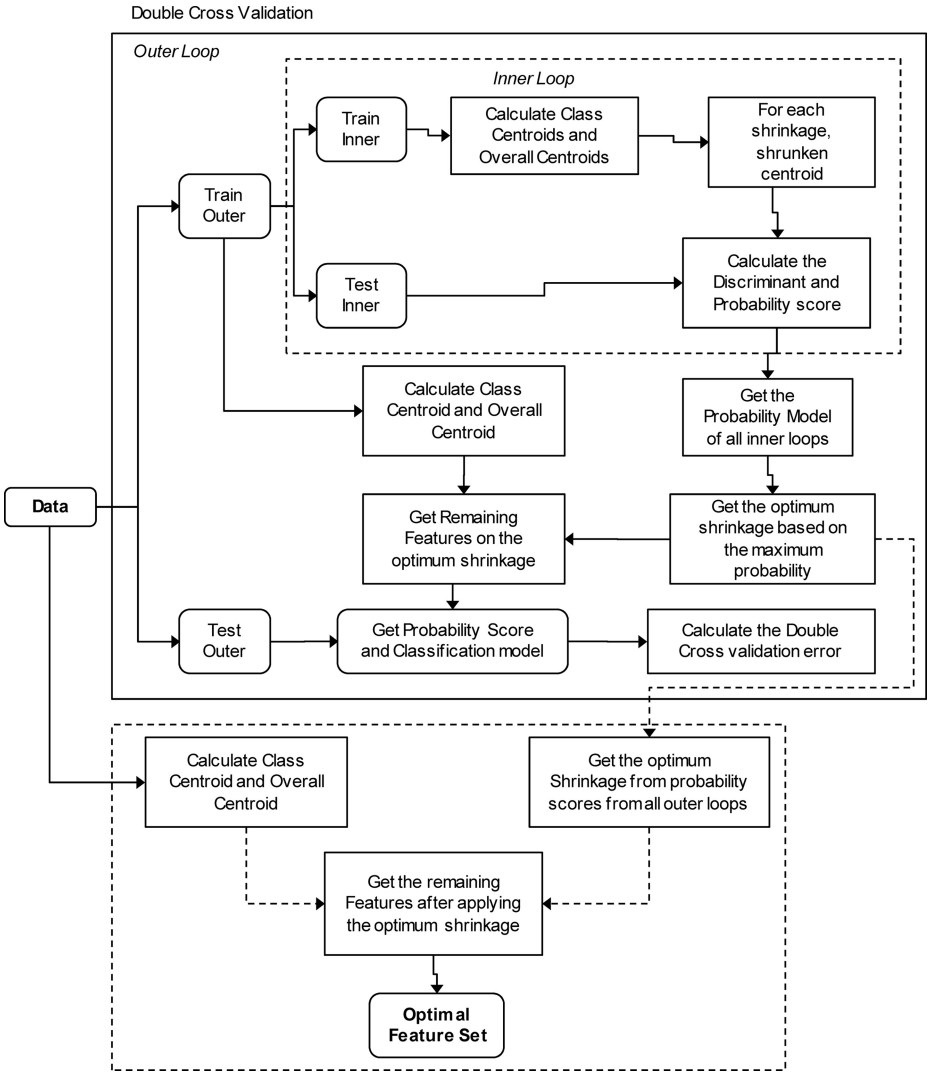

FIG. 1. **Double cross validation scheme for the nearest shrunken centroid algorithm.** In the inner loops of the double cross validation scheme, the sum of the true class probability score at the respective shrinkage is calculated (maximum provides the optimal shrinkage). The final optimal feature set is selected using the shrinkage at the maximum of the median of the sum of the true class probability and the shrinkage plot after the double cross validation procedure. Performance of the classification is measured using optimal parameters in the outer loop by calculating the classification error rate on the outer loop training data set (double cross validation error).

that constructs a hyperplane that separates two groups in a given data set. Optimal separation between two classes is reflected by obtaining a hyperplane that has the greatest distance to the nearest training data point of any class. A sample is viewed as an m-dimensional vector, where m is the number of features. The goal of the SVM is to find a hyperplane with dimension m-1 that separates the vectors based on their respective classes. The hyperplane acts as a discriminant that assigns new data to a given class. SVMs have been widely used and gained popularity for classification and prediction problems in medical research in which the feature size far exceeds the available number of samples, such as in microarray or MS analyses (23–32). Lately, approaches have been developed to adapt SVM for feature selection purposes (17, 33, 34). In this study we used a linear SVM classifier combined with a recursive feature elimination (RFE) approach for a feature selection method as introduced by Guyon *et al.* (17). This approach utilizes the weight vector **w**, which corresponds to the weight magnitude of features, as the selection criterion during RFE. The SVM-RFE procedure works as follows:

1. Initially, using all the features in the training set, train the SVM classifier.
2. Compute weight vector **w**.
3. Remove the feature with the lowest weight from the classification procedure.
4. Train the SVM classifier using the remaining features.
5. Repeat steps 2–4 until there is no remaining feature.

To obtain the optimal feature subset, we used a double cross validated SVM-RFE. The optimal number of features is determined in the inner loop. Each time the feature with the lowest weight is eliminated, the classification error based on the new set of features is calculated. Each inner loop delivers a classification error for a given set of features, and the rank of each feature is given by its weight. The optimal number of features is the smallest feature set that gives the minimum mean classification error. To select the optimum feature subset, a rank product procedure is applied to the feature rank lists produced in the inner loops. In the outer loop, the classification error of the optimal feature subset is computed using an independent test data set. The exact cross validation scheme is shown in supplemental Fig. S1.

Multivariate PCDA and PLSDA—PCDA and PLSDA take the relation between features into account in constructing new feature sets. Principal component analysis (PCA) constructs new features by finding linear transformations that best explain the variance in the data. PCA has been combined with linear discriminant analysis as the classifier applied to the PCA scores. This approach, originally proposed by Hoogerbrugge et al. (19), has been used for feature selec-

TABLE II
Confusion table; the columns correspond to features as predicted by a given method, while the rows correspond to the actual class of the features

Truth/by Methods	Selected as optimal features	Not selected
Spiked peptide-related features	True Positive (tp)	False Negative (fn)
Non-spiked peptide-related features	False Positive (fp)	True Negative (tn)

TABLE III
Definition of the scores that were used to compare the performance of different feature selection methods

Measure	Equation
Sensitivity = Recall = True Positive Rate (TPR)	$\frac{tp}{tp + fn}$
Precision	$\frac{tp}{tp + fp}$
Specificity = True Negative Rate (TNR)	$\frac{tn}{tn + fp}$
Geometric Mean Accuracy (*g*-score)	$\sqrt{TPR \cdot TNR}$
f-score	$\frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$

tion and classification in biomedical research using MS (35) or nuclear magnetic resonance data (36, 37).

PLSDA is a popular method in metabolomic studies (38–44) and has been shown to be suitable for classification and discrimination in other applications (18, 45, 46). It consists of a classical PLS regression analysis in which the response regressor is the class label. PLS components are built by trying to find a proper compromise between describing the data set and predicting the response. Further explanation and extensive assessment of this method can be found in the work of Westerhuis *et al.* (47–49).

In our study we used a combination of double and single cross validation procedures for both PCDA and PLSDA. In double cross validation, the number of principal components (PCs) (for PCDA) and the number of partial least squares (PLS) components (for PLSDA) are optimized in the inner loop. At the end of each inner loop, a rank product procedure (20) is used to rank the features based on their discriminant coefficients obtained in the inner loop. The outer loop calculates the classification error from a model that uses the optimal number of PCs/PLS components obtained in the inner loop and different numbers of features based on the ranked feature list. The feature size that gives the minimum classification error in the outer loops is selected as the optimal number of features. Based on this double cross validation procedure, the optimal number of PCs/PLS components and the optimal number of features are selected. To select the optimum feature sets, we utilized single cross validation separately from the double cross validation procedure. In the cross validation loop, the model using the optimal number of components is built, and the rank product of the discriminant coefficients of the features is calculated at the end using ranks obtained in each loop. The optimal feature set is selected from this ranked feature list, the size of which is given by the preceding double cross validation procedure. The complete scheme of PCDA and PLSDA in selecting the optimal feature set is shown in supplemental Fig. S2.

Evaluation Criteria—We performed the evaluation of the described approaches based on their performance as biomarker selection methods, rather than as learning algorithms, by measuring each algorithm's ability to construct an optimal feature set. In our case, where the discriminating features in the data sets were known, the optimal feature sets were supposed to contain only features related to the spiked peptides (true positives). True positives, false positives, true negatives, and false negatives were subsequently identified in each feature set as proposed by a given method constructing a confusion table (Table II). Several measures were calculated in order to compare and reveal the characteristics of the algorithms' performance (Table III). "Recall" expresses the proportion of selected spiked-compound-related features relative to all features that are related to the spiked peptides. "Precision" refers to the proportion of features that are related to the spiked peptides relative to all features selected by a given statistical method. The geometric mean accuracy (*g*-score) measures the ability of a method to classify both negative (not related to the spiked peptides) features and positive (spiked-peptide-related) features correctly. It assesses the overall performance of the feature selection methods, as it attributes the same importance to both true positive and true negative features. The *f*-score is a composite measure that concentrates on the correct classification of true positive features based on recall and precision. "Recall" calculates the proportion of the spiked-peptide-related features that were selected as part of the optimal feature set relative to all spiked-peptide-related features and assesses the effectiveness of an algorithm in identifying the true positive features. "Precision" is the proportion of the spiked-peptide-related features among the selected feature sets and assesses the predictive power of a method. Recall and precision are balanced in the *f*-score when the β constant parameter is set to 1 and is in favor of precision when $\beta > 1$. In our work, we set β equal to 1. We used the balanced *f*-score because we were interested in the correct identification of all spiked-peptide-related features, which requires taking both recall and precision into account to the same extent.

In addition to evaluating the optimal feature sets based on the aforementioned scores, the performance of the methods can be measured based on several additional criteria: the classification error of the learning model that is built on its selected features, the complexity/number of the selected features, and the stability of the selected feature subsets. Because the complexity of the learning model depends on the complexity of the feature set, the selected feature size has an impact on both performance and interpretability of the final model.

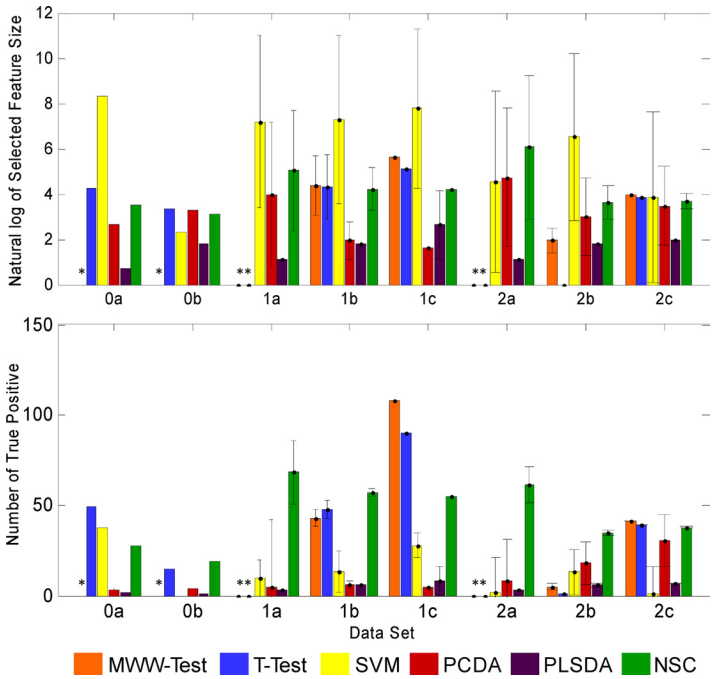

FIG. 2. **Bar charts of the median (±IQR) of the number of selected features (top) and the number of true positives (bottom) for each combination of feature selection methods and data sets (see Table I for details concerning data sets).** Results for the univariate tests (*t* test and *mww* test) on data sets 1a and 2a are denoted by **, and results of the univariate *t* test on data sets 0a and 0b are denoted by *, because these methods selected no features at a sample size of 6 or 5 respectively. Univariate *t* tests on data sets 0a and 0b (5 samples/class) and univariate tests (*t* test and *mww* test) on data sets 1c and 2c (6 samples/class) were performed once including all available samples per class without repetition.

In biomarker discovery research, the size of the feature set determines the scale of subsequent experiments, such as the identification of selected peptides or proteins and their validation as biomarkers. Thus, minimizing the number of false positives (maximizing precision) is more favorable than maximizing recall. In this study we used designed data sets in which the true positives were known. Therefore, recall and precision could be calculated and compared across different algorithms. The confidence in selecting a set of biomarker candidates can be judged by the stability of the selected feature set upon repetition of the selection procedure. More confidence is achieved when the feature selection method gives similar feature sets across multiple repetitions of cross validation runs using different sample sets. Though the stability of the obtained feature sets cannot override the classification error with respect to new test samples, it is still a useful additional criterion for selecting an optimal feature subset from different models when the list of spiked-peptide-related features is unknown.

RESULTS

Six different statistical approaches were evaluated in LC-MS data sets from human urine and porcine CSF samples spiked with a range of peptides at different concentration levels as a simulation of biomarker discovery experiments. Figs. 2–4 show the bar charts of the medians of the scores with the respective inter-quartile ranges (IQRs) from 100 repetitions for each combination of a statistical biomarker candidate selection method and a data set. The median was used because it gives robust measurements even when the distribution of the scores is not normal. Fig. 2 shows the natural logs of the feature size (top) and the number of true positives (bottom) that are contained in the corresponding feature set. Fig. 3 shows the recall (top) and precision (bottom) based on the number of true positive features found in the respective feature set. Fig. 4 shows the *f*-scores and *g*-scores, which are a composite measure of recall, precision, and true negative rate. Figs. 3 and 4 include scatter plots of median scores with error bars of recall (*y*-axis) and precision (*x*-axis) and *g*- (*y*-axis) and *f*-scores (*x*-axis), respectively. These scores were used to compare and assess the performance of each method with respect to sample size, within- and between-class variability of spiked peptides, type of biological background, and type of mass spectrometer.

Comparison of Individual Methods—Performance assessment of the various feature selection methods starts with discussions of results obtained with data sets 0a and 0b. These two data sets are composed of two different biological backgrounds (human urine for 0a and porcine CSF for 0b), a sample size of 5, and a large between-class variability of spiked peptides. Both classes are composed of samples of one spiking level, resulting in data sets with low within-class variability of spiked peptides. In these data sets, identification of biomarkers represented here as spiked-in compounds is relatively simple. This is followed by a detailed discussion of the more challenging data sets 1a–1c and 2a–2c, which are composed of human urine samples of different sample sizes (6, 12, and 15 corresponding to a, b,

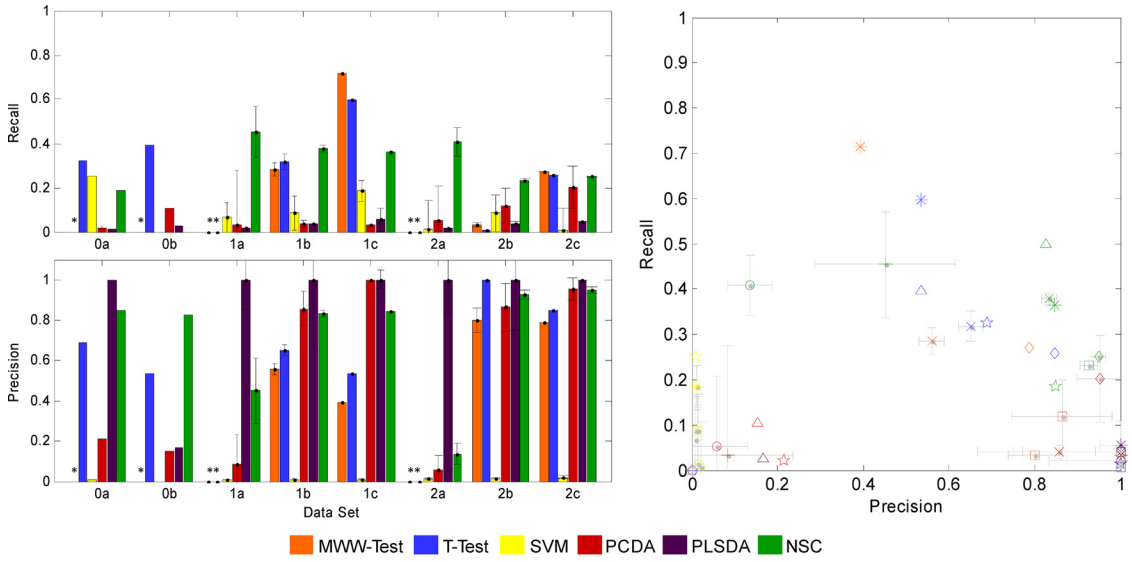

FIG. 3. **Bar charts (left) and scatter plot (right) of the median (±IQR) of recall (top) and precision (bottom) for each combination of feature selection methods and data sets.** Recall and precision were not available for the *t* test and the *mww* test on data sets 1a and 2a, containing 6 samples (denoted by **), or for the *mww* test on data sets 0a and 0b, containing 5 samples (denoted by *). Univariate *t* tests on data sets 0a and 0b (5 samples/class) and univariate tests (*t* test and *mww* test) on data sets 1c and 2c (6 samples/class) were performed once including all available samples per class without repetition. Gray error bars in the scatter plot show the IQRs of recall and precision.

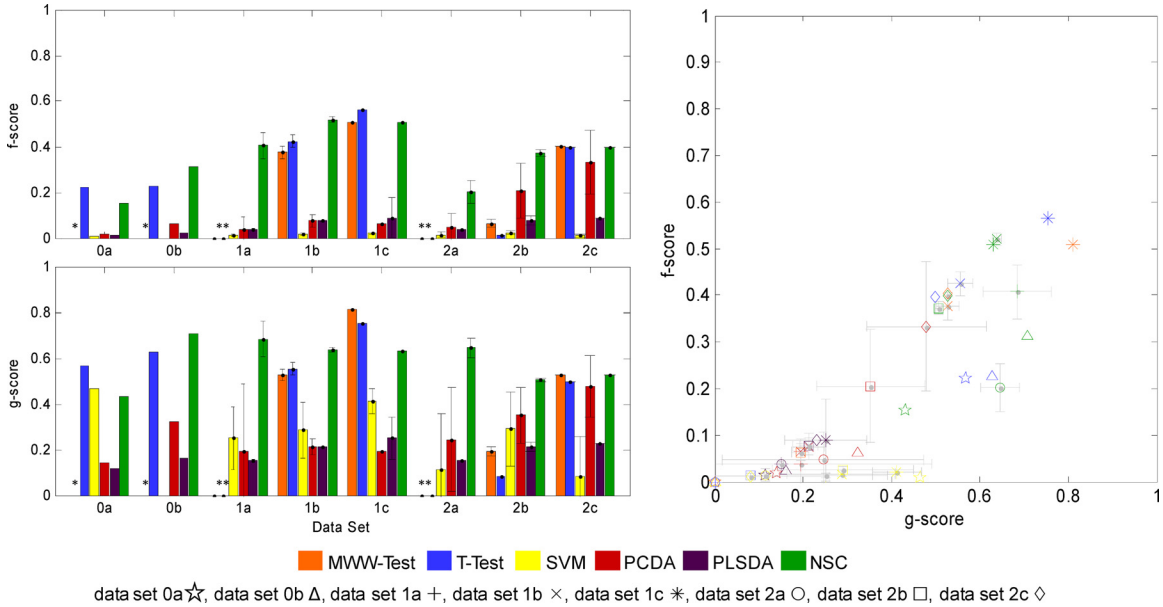

FIG. 4. **Bar charts (left) and scatter plot (right) of the median (±IQR) *f*-score (top) and *g*-score (bottom) for each combination of feature selection methods and data sets.** The *f*-score and *g*-score were not available for the *t* test and the *mww* test on data sets containing 6 samples (denoted by **) or for the *mww* test on data sets 0a and 0b containing 5 samples (denoted by *). Univariate *t* tests on data sets 0a and 0b (5 samples/class) and univariate tests (*t* test and *mww* test) on data sets 1c and 2c (6 samples/class) were performed once including all available samples per class without repetition. Gray error bars in the scatter plot show the IQRs of recall and precision.

and c in our data set notation) and low (8×) and large (32×) between-class variability of spiked peptides (corresponding to 1 and 2 in the data set notation). Finding spiked-in compounds in these data sets is more challenging, as the within-class variability of spiked peptides is much broader (4×) than the small within-class variability corresponding to analytical variance in data sets 0a and 0b. The larger within-class variance of spiked peptides relative to what is expected from analytical variance in these data sets mimics the expected behavior of biomarkers better than high and low classes with fixed spiking levels, because potential biomarkers may be present in sample groups with a large concentration distribu-

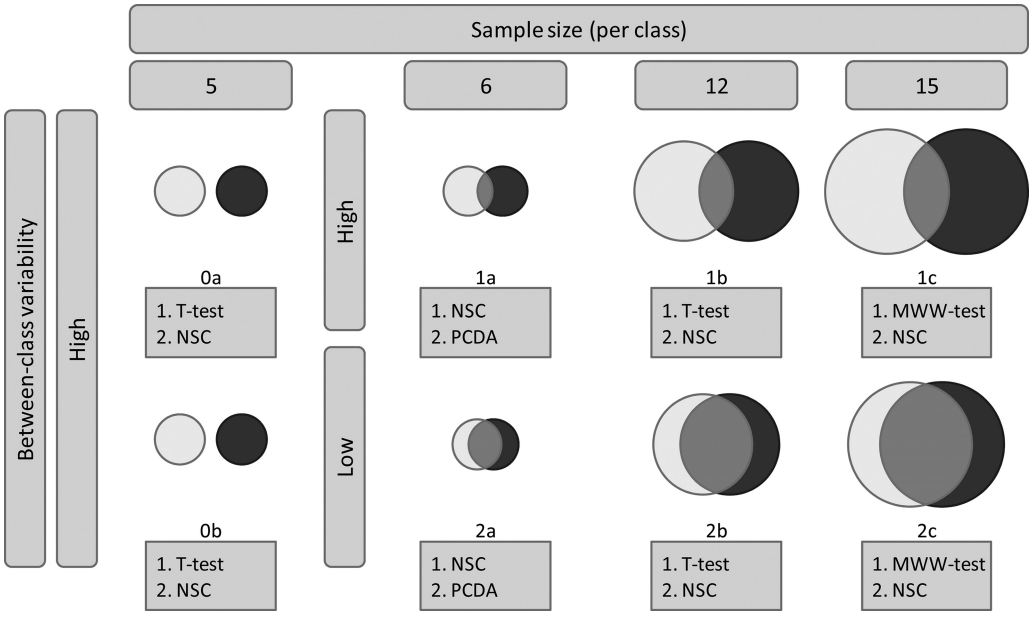

FIG. 5. **Overview of the two best performing feature selection statistical methods for data sets of different sample sizes and between- and within-class variability of spiked peptides based on the *f*-score.** NSC shows the best performance for data sets with 6 samples independent of between- and within-class variability of spiked peptides, whereas univariate tests rank on top when the sample size increases to 15 samples per class or for low-sample-size data sets (0a and 0b) with low within-class variability of spiked peptides.

tion. In addition, compounds may be present in only a fraction of the samples in sample groups where the within-class concentration distribution of spiked compound crosses the limit of detection. Table I and Fig. 5 contain additional information on the data set design.

t Test and mww Test—Figs. 3 and 4 show that for data sets 0a and 0b, the *t* test provides high precision and *g*-scores greater than 0.5, resulting in greater precision than with recall. For these data sets, the *t* test provides the most complete list of true positives, with fewer false negatives and false positives than the other feature selection methods. The biological background and the type of mass spectrometer have little influence on the performance of the *t* test as indicated by the results of data sets 0a and 0b. For these data sets, the *mww* test cannot be applied because of the low sample size; rank-based *mww* tests would result in tied ranks and tied *p* values.

Figs. 2 and 3 show that neither the *t* test nor the *mww* test is capable of finding discriminating features in the two classes in the more challenging data sets with large within-class variances of spiked-in compounds when the sample size is low (data sets 1a and 2a, six samples per class), independent of the magnitude of the between-class variance of spiked peptides. This improves markedly when the number of samples per class is increased to 15 (data sets 1c and 2c), at which point both tests are among the best performing methods. As expected from univariate methods, the sample size has a strong influence on performance, with a clear threshold between no or very poor performance for 6 samples per class and rather good performance for 15 samples per group. Similar results were obtained with data sets 0a and 0b (with a sample size of 5) and data sets 1b and 1c (with sample sizes of 12 and 15), indicating that *t* test performance is influenced by the within-class variance of the biomarker candidates. When the within-class variance is low, a lower sample size is sufficient to reach adequate performance of true positive selection with a low number of false positives and false negatives. Fig. 3 shows, however, that the high number of true positives in the results from data sets with 15 samples and a large between-class variance of spiked compounds (see Fig. 2, data set 1c) are accompanied by a relatively high number of false positives, lowering precision despite a high recall. Despite modest precision, univariate feature selection methods gave the highest *g*-scores and *f*-scores for these data sets (Fig. 4, data sets 0a, 0b, and 1c), showing that the trade-off between recall and precision results overall in the best performance with respect to *g*- and *f*-scores. It is interesting to note that univariate tests gave a lower recall but a higher precision when between-class variance decreased (Fig. 3, data set 2c), which is primarily attributable to a 2-fold lower number of detected true positives (Fig. 2, data set 2c). The same tendency was observed for data sets 1b and 2b, with 12 samples per class. The composite *f*-score and *g*-score show that the overall performance of univariate feature selection methods is mainly affected by the sample size and within-class variability, and slightly by the between-class variability, of spiked-in peptides. It shows as well that standard univariate feature selection methods perform as well as or even better than the more sophisticated multivariate or semi-multivariate methods with a class size of 15 samples in the case of large within-class variability of spiked-in peptides, or with a

smaller sample size of 5 when the within-class variability of spiked-in peptides is negligible.

NSC—The use of NSC resulted in the highest number of true positive features with a relatively low number of false positive features in data sets with low within-class variability (data sets 0a and 0b) regardless of the biological background (urine or CSF) or the type of mass spectrometer (low or high resolution), leading to high precision. However, NSC did not provide a complete list of true positive features, resulting in low recall (0.185 in the case of human urine in data set 0a, and 0.50 for CSF in data set 0b). Similarly, this algorithm provided the highest number of true positives for the smallest sample size of 6 samples per class with large within-class variance of spiked-in peptides (data sets 1a and 2a), but at the expense of selecting a fairly high number of features that were not related to the spiked peptides (Fig. 2, data sets 1a and 2a). This leads to high recall but intermediate precision relative to the other algorithms (Fig. 3, data sets 1a and 2a). Notably, PLSDA outperformed NSC with respect to precision for data sets with a low number of samples because of the very low number of selected features in data sets with large within-class variability of spiked-in compounds (data sets 1a and 2a) that are not related to the spiked peptides (false positives). NSC remains among the better performing algorithms for data sets with a higher number of samples (data sets 1b, 1c, 2b, and 2c), showing fairly stable performance across all evaluated data sets, which is also reflected in rather similar *g*-scores and *f*-scores (Fig. 4). NSC benefits from increasing sample size when it comes to precision (Fig. 3), as it selects fewer features that are not related to the spiked peptides, whereas the number of true positives decreases only slightly, resulting in improved precision without a significant sacrifice of recall. It is also noteworthy that the robustness of the statistical model improved with increasing sample size based on the reduced IQR (see error bars in Figs. 2–4). The difference in recall and precision between data sets with 12 and 15 samples was not significant (Fig. 3, data set 1b *versus* 1c, and 2b *versus* 2c). Higher recall and slightly lower precision were observed for data sets with large class separation (1b and 1c) relative to those with small class separation (Fig. 3, data sets 1b and 1c *versus* data sets 2b and 2c), which holds also for both the *f*-score and the *g*-score (Fig. 4).

SVM-RFE—SVM-RFE selected the highest number of features in almost all data sets, whereas the number of selected true positives was lower than for most of the other methods, resulting in many false positives (Fig. 2). There were large differences between the number of selected features in data sets with low within-class variability of spiked-in peptides (4100 and 10 selected features in data sets 0a and 0b, respectively). This large difference may be explained by the poor performance of SVM-RFE, as no true positive was included in the low number of selected features in the porcine CSF data set 0b, and only 38 true positives were included in the selected 4100 features from the human urine sample. The results of both data sets seem therefore to be a random selection of features without any preference to selectively enrich true positives. Even though there is a trend toward better performance as the sample size and between-class variability of spiked peptides increase (data set 1), scores remain low, with a maximum recall of 0.2 and a maximum precision of 0.05 (Fig. 3). The large number of selected features lowers precision and, consequently, the *f*-score (Figs. 3 and 4). The *g*-score is also lower than for most other algorithms. It is a property of SVM that many correlated variables receive almost equal weights, which means that the weight of a feature is not a very useful measure for feature selection. This could be why the size of the feature set selected by SVM-RFE is rather large. Our results with peptide-spiked human urine or porcine CSF samples show that the SVM-RFE approach is not suitable for the selection of biomarker candidates.

PCDA—For most of the studied data sets, PCDA tended to select a low number of true positive features and a low number of features resulting in low recall and high precision, notably for large sample sizes, although results fluctuated considerably (Figs. 2 and 3, data sets 1b, 1c, 2b, and 2c). PCDA may thus be considered a fairly "conservative" approach to biomarker discovery, as the selected feature list has a relatively high content of true positive features. In contrast to NSC and the univariate tests (*t* test and *mww* test), PCDA tends to select fewer features from data sets with high between-class variability (data set 1); this is most pronounced for data set 1c. While the number of selected features is low (mean of 5.3 $\pm$ 0.8), the selection contains essentially only true positive features (mean of 5.13 $\pm$ 0.8), resulting in very high precision (mean of 0.97 $\pm$ 0.06) in data sets 1a–1c and 2a–2c. This may be considered a desirable characteristic of this approach when it comes to subsequent validation of the selected biomarker candidates in large numbers of samples. However, most of the true positives are missed using this statistical approach. Accurate selection of a low number of true positive features with low recall is reflected by the poor values of the composite measure *f*-scores and *g*-scores (Fig. 4). The precision of the PCDA method is lower in data sets with low sample sizes irrespective of the between- and within-class variability of spiked-in peptides (Fig. 3, data sets 0a, 0b, 1a, and 2a), which makes the method adequate for biomarker candidate selection in data sets having a sample size equal to or greater than 12 samples per group. Similar results for data sets 0a, 0b, 1a, and 2a show that the performance of PCDA is not affected by the biological background, resolution of the mass spectrometer, or within-class variability of spiked peptides.

PLSDA—The most striking characteristic of PLSDA is the extremely high precision regardless of the sample size or the class separation (Fig. 3) in almost all data sets of spiked human urine samples. Surprisingly, the analysis of spiked trypsin-digested porcine CSF yielded lower precision due to a

relatively low number of true positives among the selected features (1 out of 6), which could be related to the lower number of spiked features (38, compared with 151 in human urine samples). The selection of a low number of true positive features with high precision is advantageous in cases when subsequent biomarker validation is tedious, requires significant effort, and resembles the results obtained with PCDA. The stability of the model underlying feature selection increases with increasing sample size, as shown by the reduced IQR. High precision comes at a price, however, as the number of selected true positives is small relative to the number of expected true positive features based on the spiked peptides (Fig. 2). The low number of selected true positives likely results from the facts that PLSDA and PCDA select with high probability the most abundant features with relative low error and that the cut-off value for the rank products for selecting features is too conservative. Because signals related to the spiked peptides are highly correlated, PCDA and PLSDA select only a few of them that represent the between-class variability of spiked peptides well. Globally, PLSDA is not really affected by the strength of the between-class variability of spiked peptides in the data set, because the patterns of recall and precision in data sets 1 and 2 are comparable (Fig. 3). When sample size increases, recall improves slightly, although it remains low overall. In contrast to PCDA, PLSDA also shows high precision on data sets with a low sample size (Fig. 3, data sets 0a, 1a, and 2a), making this method more adequate for accurately selecting true positive features in data sets with only 6 samples per group than PCDA.

Comparison between Methods—Although all methods benefit from a larger sample size, only some of them are affected by the between- and within-class variability of spiked peptides. The univariate *t* test and *mww* test results are strongly affected by the between- and within-class variability of spiked peptides (based on the comparison of *f*-score, *g*-score, recall, and precision). Furthermore, they require a minimum sample size to function. Multivariate methods that use feature transformation prior to selecting a given feature set, such as PCDA and PLSDA, are not strongly affected by the between-class variability of spiked peptides or sample size. The performance of NSC is overall rather independent of the between-class variability of spiked peptides and sample size.

When the characteristics of the data set were profitable (large sample size and high between-class variability of spiked peptides), univariate methods (*t* test and *mww* test) performed best, because they assigned most of the true positives within a reasonably sized total feature set. They were furthermore the fastest and simplest methods to use. Univariate methods failed when the sample size was small (*e.g.* six samples per class), except when the between-class variability of spiked peptides was high and the within-class variability of spiked peptides was low. This latter condition is, however, very unlikely in biomarker discovery, in which compounds have large within-class variability due to large biological variability and tend to be present to some extent in all sample classes. Based on the *f*-score, *g*-score, recall, and precision, the semi-multivariate NSC outperformed all other methods, including multivariate methods, in terms of feature selection, as it strikes the best balance between recall and precision, keeping both *f*-scores and *g*-scores high. The multivariate methods PLSDA and PCDA provided high-quality feature sets that are reflected in high precision approaching 100% at the expense of a low recall. Globally speaking, NSC is applicable to all tested data sets and might be considered a good compromise when performing small-scale biomarker discovery studies.

Additional assessment criteria are the classification error rate or sum of true class probability in the case of NSC and the stability of the models. When repeating the calculation a number of times, the result is more trustworthy if all repetitions yield similar conclusions. To test this, we assessed the variability of feature selection performance across 100 repetitions in data sets with sample sizes of 12 and 15 (1b, 1c, 2b, and 2c) based on *g*-score and *f*-score, variation of the classification error rate, and the sum of true class probability in the case of NSC. The variability of *g*-scores and *f*-scores was measured using the IQR for a given data set. In general, there was a tendency for the IQR to decrease with all approaches as the number of samples increased (see error bars in Figs. 2–4), except for SVM-RFE, which might be due to the poor overall performance of the approach resulting in quasi-randomly selected feature sets.

The variability of the classification error rate or the sum of true class probability in each cross validation loop reflects the stability of the model that is used as a classifier. Two different kinds of classification error can be derived from such a double cross validation scheme: the error in the inner loops, which determines the optimal values for the parameters, and the classification error in the outer loops when using optimal parameter values, which assesses the error rate of the optimal model for classifying new sets of samples. The inner loop classification error rate or sum of true class probability shows considerable dependence on the sample size of the data sets, as shown in supplemental Figs. S3–S6. However, the between-class variability of spiked peptides does not have an effect on inner loop classification error rates or on the sum of true class probabilities. The error in the outer loop averaged 20% for data sets with high between-class variability of spiked peptides (1a–1c) and 30% for data sets with low between-class variability of spiked peptides (2a–2c) independent of the applied method. The fluctuation of the inner loop classification error rate or the sum of true class probability is due to different values for the optimal parameters determined in the inner loops obtained from varying training samples sets and can be assessed in a classification error or sum of true class probability against a parameter plot. These

plots indicate whether the minimum error, which determines the optimum parameter value, is located in a smooth/stable region. All plots that were used to determine the optimal parameters in this study are shown in supplemental Figs. S3–S6. The spread of the error decreased with increasing sample size, showing that the stability of the models increased with increasing sample number. The results show also that it is not justifiable to rely on a classification result from a model obtained from a training data set with only six samples per group, except in the case of the NSC algorithm.

To assess the variability of the selected feature sets delivered by each method, we compared the count of features that were selected at least once across 100 repetitions (unique features) relative to the count of features that were selected in each repetition (common features) (shown in supplemental Table S2). The stability of the feature set is not available because only one repetition was possible for all feature selection methods for data sets 0a and 0b because of the low sample size. Similarly, the stability of the feature set was not determined for the *mww* test or the *t* test for data sets 1c and 2c because of the lack of a double cross validation scheme for these methods. From this result, NSC produced the most stable feature set of all methods, as shown by the high ratio of the number of common features to the number of unique features.

DISCUSSION

We have assessed different feature selection methods with respect to their capacity to deliver biomarker candidates from a number of well-controlled data sets that were obtained via LC-MS analysis of peptide-spiked human urine or trypsin-digested porcine CSF samples. Six widely used feature selection methods were compared and their performance measured based on how well they found true positives (features that are related to the spiked peptides) and how well they avoided false positives (all other features) for data sets with different sample sizes and between- and within-class variability of spiked peptides. We derive six main conclusions from this study. (1) As expected, all methods benefit from a larger sample size. (2) Univariate methods and semi-multivariate methods are more sensitive to the between- and within-class variability of spiked peptides, whereas multivariate methods (especially PLSDA) are hardly affected by the between-class variability of spiked peptides. (3) SVM-RFE performed poorly on all data sets with respect to selecting relevant features, showing that the weight vector is not a suitable criterion for feature selection/elimination. (4) True multivariate methods like PCDA and PLSDA aim at high precision by sacrificing recall (*i.e.* they are conservative with respect to selecting true positive features). PLSDA performs better than PCDA for data sets with a low sample size when the within-class variability of spiked peptides is high. (5) The semi-multivariate NSC strikes the best compromise between recall and precision regardless of sample size and between- and within-class variability of spiked peptides. (6) The performance of feature selection methods shows little dependence on the number of discriminating features, the biological background, or the type of mass spectrometer used for data acquisition. Fig. 5 provides an overview of the best performing feature selection methods based on the *f*-score for data sets with different within- and between-class variability of spiked peptides and sample size. The figure provides a summary concerning the choice of a given feature selection method and should help practitioners select the most suitable method for biomarker discovery studies.

Because biomarker discovery is usually intended to support clinical diagnosis, it is advantageous to obtain a discriminating feature set with a minimum number of false positives and with the potential to classify new sets of samples correctly. Based on this criterion, PLSDA is a good choice because of its excellent precision. When additional, correlated discriminating features are required—for example, to support pathway analysis—PLSDA might miss relevant features that could be informative. In that case, NSC provides a better compromise between recall and precision, with a higher number of true positives at a reasonable false positive rate. In cases in which data from more samples are available (more than 15 samples per group/class in our case), univariate tests (*t* test or *mww* test with multiple testing correction) are able to identify biomarker candidates with high confidence. For classes with low sample numbers (six samples per class in our case), NSC has the greatest potential to select biomarker candidates successfully. However, our results show that there is considerable danger in relying on results from data sets with such a small sample size, as classification models and the values for optimized parameters are prone to significant fluctuations, making biomarker selection uncertain.

To enhance the application of the presented feature selection methods and further the use of our test data sets to assess other feature selection algorithms, we have made the source code and preprocessed LC-MS data with descriptions of spiking levels and spiked compounds and indices of spiked-in-peptide-related features available through the source code repository of Netherlands Bioinformatics Centre (link is on page 2 in supplementary material).

* The work described in this publication was supported by grant BioRange 2.2.3 from the Netherlands Bioinformatics Center and from joint Gaining Momentum Initiative of the Netherlands Bioinformatics and the Netherlands Proteomics Centers.

[S] This article contains supplemental material.

** To whom correspondence should be addressed. Tel.: +31-50-363-3341; Fax: +31-50-363-7582; E-mail: p.l.horvatovich@rug.nl.

REFERENCES

1. Mischak, H., Allmaier, G., Apweiler, R., Attwood, T., Baumann, M., Benigni, A., Bennett, S. E., Bischoff, R., Bongcam-Rudloff, E., Capasso, G., Coon, J. J., D'Haese, P., Dominiczak, A. F., Dakna, M., Dihazi, H., Ehrich, J. H., Fernandez-Llama, P., Fliser, D., Frokiaer, J., Garin, J., Girolami, M., Hancock, W. S., Haubitz, M., Hochstrasser, D., Holman, R. R., Ioannidis, J. P., Jankowski, J., Julian, B. A., Klein, J. B., Kolch, W., Luider, T., Massy, Z., Mattes, W. B., Molina, F., Monsarrat, B., Novak, J., Peter, K.,

Rossing, P., Sanchez-Carbayo, M., Schanstra, J. P., Semmes, O. J., Spasovski, G., Theodorescu, D., Thongboonkerd, V., Vanholder, R., Veenstra, T. D., Weissinger, E., Yamamoto, T., and Vlahou, A. (2010) Recommendations for biomarker identification and qualification in clinical proteomics. *Sci. Transl. Med.* **2,** 46ps42
2. Puntmann, V. O. (2009) How-to guide on biomarkers: biomarker definitions, validation and applications with examples from cardiovascular disease. *Postgrad. Med. J.* **85,** 538–545
3. Rifai, N., Gillette, M. A., and Carr, S. A. (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **24,** 971–983
4. Hoekman, B., Breitling, R., Suits, F., Bischoff, R., and Horvatovich, P. (2012) msCompare: a framework for quantitative analysis of label-free LC-MS data for comparative candidate biomarker studies. *Mol. Cell. Proteomics* **11,** M111.015974
5. Saeys, Y., Inza, I., and Larranaga, P. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* **23,** 2507–2517
6. Smit, S., Hoefsloot, H. C. J., and Smilde, A. K. (2008) Statistical data processing in clinical proteomics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **866,** 77–88
7. Smit, S., van Breemen, M. J., Hoefsloot, H. C., Smilde, A. K., Aerts, J. M., and de Koster, C. G. (2007) Assessing the statistical validity of proteomics based biomarkers. *Anal. Chim. Acta* **592,** 210–217
8. Kohavi, R., and John, G. H. (1997) Wrappers for feature subset selection. *Artif. Intell.* **97,** 273–324
9. Hilario, M., and Kalousis, A. (2008) Approaches to dimensionality reduction in proteomic biomarker studies. *Brief Bioinform.* **9,** 102–118
10. Baek, S., Tsai, C. A., and Chen, J. J. (2009) Development of biomarker classifiers from high-dimensional data. *Brief Bioinform.* **10,** 537–546
11. Datta, S., and Pihur, V. (2010) Feature selection and machine learning with mass spectrometry data. *Methods Mol. Biol.* **593,** 205–229
12. Van der Walt, C., and Barnard, E. (2006) Data characteristics that determine classifier performance. SAIEE Africa Research Journal, **98,** 87–93
13. Rubingh, C., Bijlsma, S., Derks, E., Bobeldijk, I., Verheij, E., Kochhar, S., and Smilde, A. (2006) Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics* **2,** 53–61
14. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300
15. Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 6567–6572
16. Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.* **18,** 104–117
17. Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46,** 389–422
18. Barker, M., and Rayens, W. (2003) Partial least squares for discrimination. *J. Chemom.* **17,** 166–173
19. Hoogerbrugge, R., Willig, S. J., and Kistemaker, P. G. (1983) Discriminant analysis by double stage principal component analysis. *Anal. Chem.* **55,** 1710–1712
20. Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* **573,** 83–92
21. Suits, F., Hoekman, B., Rosenling, T., Bischoff, R., and Horvatovich, P. (2011) Threshold-avoiding proteomics pipeline. *Anal. Chem.* **83,** 7786–7794
22. Vapnik, V. (1998) *Statistical Learning Theory*, Wiley-Interscience
23. Anderson, D. C., Li, W., Payan, D. G., and Noble, W. S. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.* **2,** 137–146
24. Mao, Y., Zhou, X. B., Pi, D. Y., and Sun, Y. X. (2005) Constructing support vector machine ensembles for cancer classification based on proteomic profiling. *Genomics Proteomics Bioinformatics* **3,** 238–241
25. Jiang, Z., Yamauchi, K., Yoshioka, K., Aoki, K., Kuroyanagi, S., Iwata, A., Yang, J., and Wang, K. (2006) Support vector machine-based feature selection for classification of liver fibrosis grade in chronic hepatitis C. *J. Med. Syst.* **30,** 389–394
26. Guo, J., Deng, W., Zhang, L., Li, C., Wu, P., and Mao, P. (2007) Prediction of prostate cancer using hair trace element concentration and support vector machine method. *Biol. Trace Elem. Res.* **116,** 257–272
27. Mao, Y., Zhao, X., Wang, S., and Cheng, Y. (2007) Urinary nucleosides based potential biomarker selection by support vector machine for bladder cancer recognition. *Anal. Chim. Acta* **598,** 34–40
28. Lin, E., and Hwang, Y. (2008) A support vector machine approach to assess drug efficacy of interferon-alpha and ribavirin combination therapy. *Mol. Diagn. Ther.* **12,** 219–223
29. Pham, T. V., van de Wiel, M. A., and Jimenez, C. R. (2008) Support vector machine approach to separate control and breast cancer serum samples. *Stat. Appl. Genet. Mol. Biol.* **7,** Article 11
30. Webb-Robertson, B. J., Cannon, W. R., Oehmen, C. S., Shah, A. R., Gurumoorthi, V., Lipton, M. S., and Waters, K. M. (2008) A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics* **24,** 1503–1509
31. Henneges, C., Bullinger, D., Fux, R., Friese, N., Seeger, H., Neubauer, H., Laufer, S., Gleiter, C. H., Schwab, M., Zell, A., and Kammerer, B. (2009) Prediction of breast cancer by profiling of urinary RNA metabolites using support vector machine-based feature selection. *BMC Cancer* **9,** 104
32. Zou, A. M., Wu, F. X., Ding, J. R., and Poirier, G. G. (2009) Quality assessment of tandem mass spectra using support vector machine (SVM). *BMC Bioinformatics* **10 Suppl 1,** S49
33. Hermes, L., and Buhmann, J. M. (2000) Feature selection for support vector machines. *Proceedings of the 15th International Conference on Pattern Recognition, 2000,* Barcelona, Spain September 3–7 pp. 712–715
34. Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. (2000) Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 27 Nov- 2 Dec 2000, Denver, CO, MIT Press 2001 Feature selection for SVMs.
35. Hoefsloot, H. C., Smit, S., and Smilde, A. K. (2008) A classification model for the Leiden proteomics competition. *Stat. Appl. Genet. Mol. Biol.* **7,** Article 8
36. Amato, U., Larobina, M., Antoniadis, A., and Alfano, B. (2003) Segmentation of magnetic resonance brain images through discriminant analysis. *J. Neurosci. Meth.* **131,** 65–74
37. Lamers, R. J., DeGroot, J., Spies-Faber, E. J., Jellema, R. H., Kraus, V. B., Verzijl, N., TeKoppele, J. M., Spijksma, G. K., Vogels, J. T., van der Greef, J., and van Nesselrooij, J. H. (2003) Identification of disease- and nutrient-related metabolic fingerprints in osteoarthritic guinea pigs. *J. Nutr.* **133,** 1776–1780
38. Ramadan, Z., Jacobs, D., Grigorov, M., and Kochhar, S. (2006) Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. *Talanta* **68,** 1683–1691
39. Lv, Y., Liu, X., Yan, S., Liang, X., Yang, Y., Dai, W., and Zhang, W. (2010) Metabolomic study of myocardial ischemia and intervention effects of Compound Danshen Tablets in rats using ultra-performance liquid chromatography/quadrupole time-of-flight mass spectrometry. *J. Pharm. Biomed. Anal.* **52,** 129–135
40. Liu, Y., Huang, R., Liu, L., Peng, J., Xiao, B., Yang, J., Miao, Z., and Huang, H. (2010) Metabonomics study of urine from Sprague-Dawley rats exposed to Huang-yao-zi using (1)H NMR spectroscopy. *J. Pharm. Biomed. Anal.* **52,** 136–141
41. Lan, K., Zhang, Y., Yang, J., and Xu, L. (2010) Simple quality assessment approach for herbal extracts using high performance liquid chromatography-UV based metabolomics platform. *J. Chromatogr. A* **1217,** 1414–1418
42. Kim, H. K., Saifullah Khan, S., Wilson, E. G., Kricun, S. D., Meissner, A., Goraler, S., Deelder, A. M., Choi, Y. H., and Verpoorte, R. (2010) Metabolic classification of South American Ilex species by NMR-based metabolomics. *Phytochemistry* **71,** 773–784
43. Feng, B., Wu, S. M., Lv, S., Liu, F., Chen, H. S., Gao, Y., Dong, F. T., and Wei, L. (2009) A novel scoring system for prognostic prediction in d-galactosamine/lipopolysaccharide-induced fulminant hepatic failure BALB/c mice. *BMC Gastroenterol.* **9,** 99
44. Barba, I., Garcia-Ramirez, M., Hernandez, C., Alonso, M. A., Masmiquel, L., Garcia-Dorado, D., and Simo, R. (2010) Metabolic fingerprints of proliferative diabetic retinopathy: an 1H-NMR-based metabonomic approach using vitreous humor. *Invest. Ophthalmol. Vis. Sci.* **51,** 4416–4421
45. Boulesteix, A. L., and Strimmer, K. (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform.*

8, 32–44
46. Chevallier, S., Bertrand, D., Kohler, A., and Courcoux, P. (2006) Application of PLS-DA in multivariate image analysis. *J. Chemom.* **20,** 221–229
47. Westerhuis, J., Hoefsloot, H., Smit, S., Vis, D., Smilde, A., van Velzen, E., van Duijnhoven, J., and van Dorsten, F. (2008) Assessment of PLSDA cross validation. *Metabolomics* **4,** 81–89
48. Westerhuis, J., van Velzen, E., Hoefsloot, H., and Smilde, A. (2008) Discriminant Q2 (DQ2) for improved discrimination in PLSDA models. *Metabolomics* **4,** 293–296
49. Westerhuis, J. A., van Velzen, E. J., Hoefsloot, H. C., and Smilde, A. K. (2010) Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics* **6,** 119–128